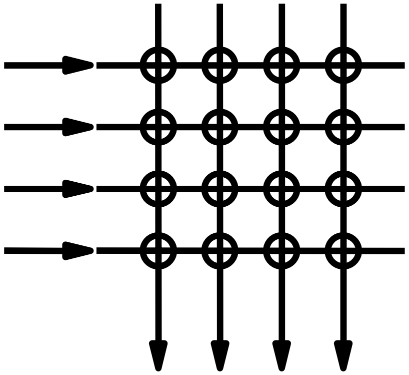
Benes Switching Fabrics with O(N)-Complexity Internal Backpressure

Georgios Sapountzis and Manolis Katevenis

FORTH & Univ. of Crete, Greece

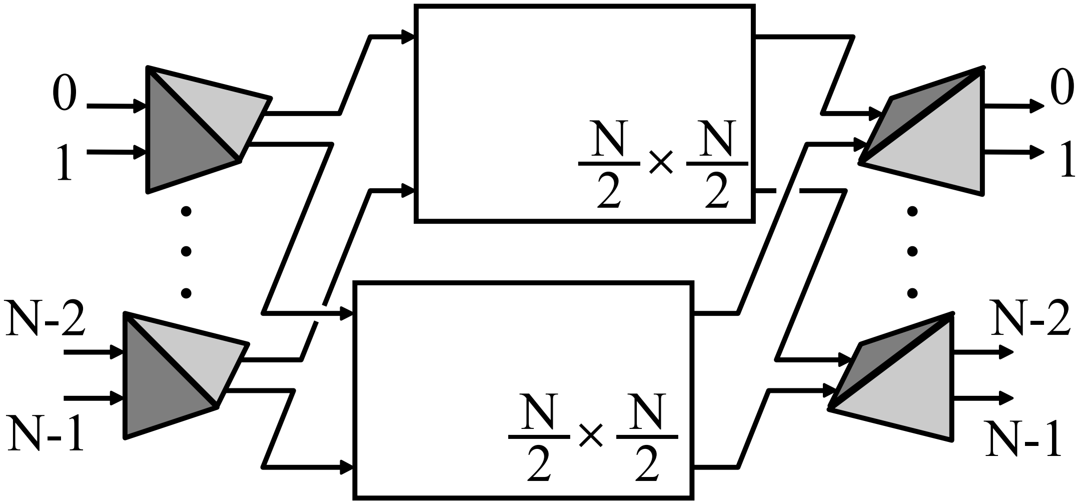
Scalable Non-Blocking Switching

- Crossbar:
  + simple and regular, but
  – O(N^2) cost.

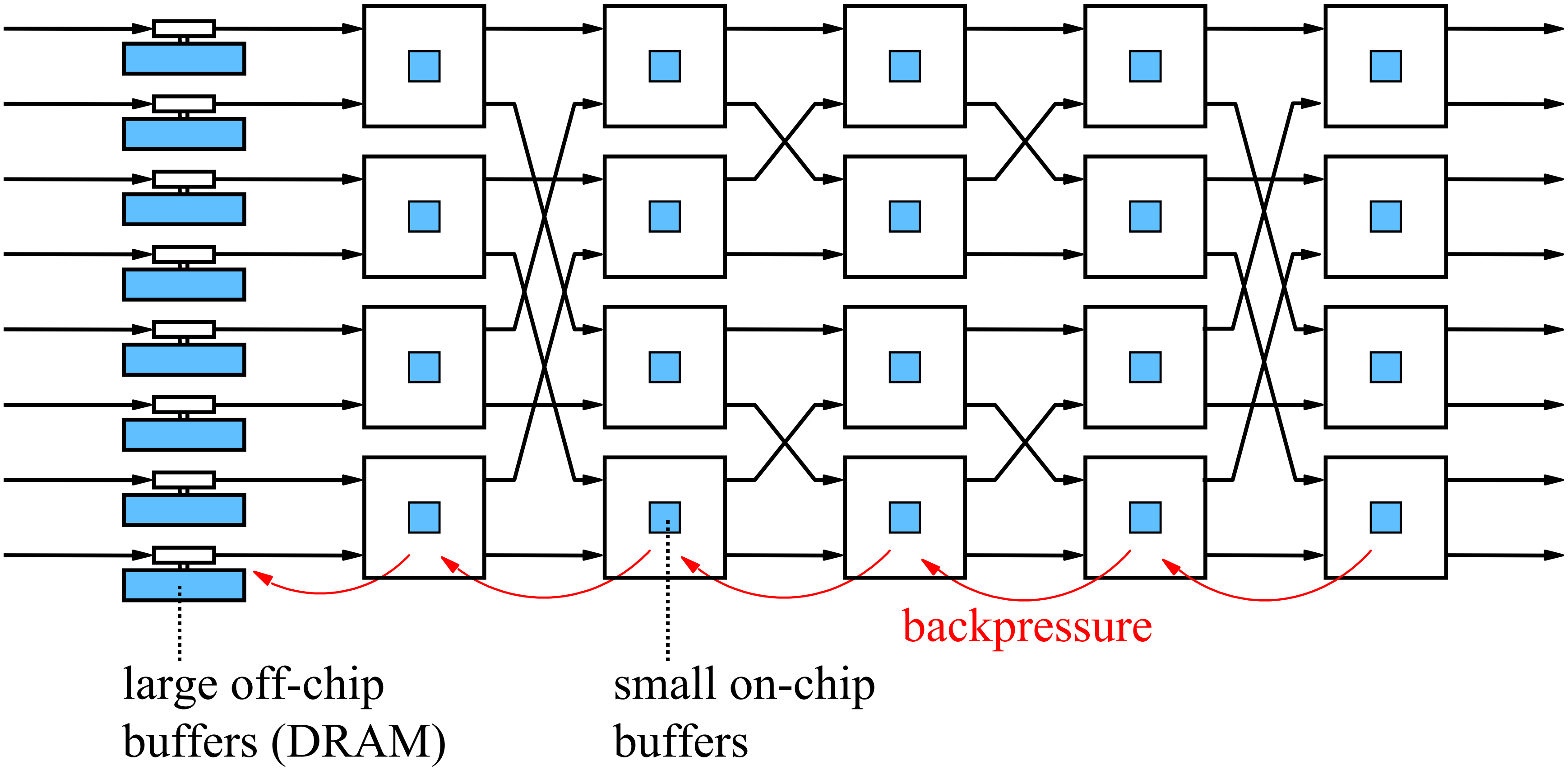


- Benes fabric:
  + O(N•log N) cost,
  + non-blocking,

  inverse multiplexing
  - multi-path routing
  - re-sequencing
  - load balancing


0
1
N-2
N-1
$\frac{N}{2} \times \frac{N}{2}$
$\frac{N}{2} \times \frac{N}{2}$
0
1
N-2
N-1

Buffered Switching Fabrics with Internal Backpressure



- Performance of OQ at the cost of IQ,
- Requires per-flow backpressure.

This Work:

- Multi-path routing & re-sequencing + per-flow backpressure.
- Flow merging to reduce cost.

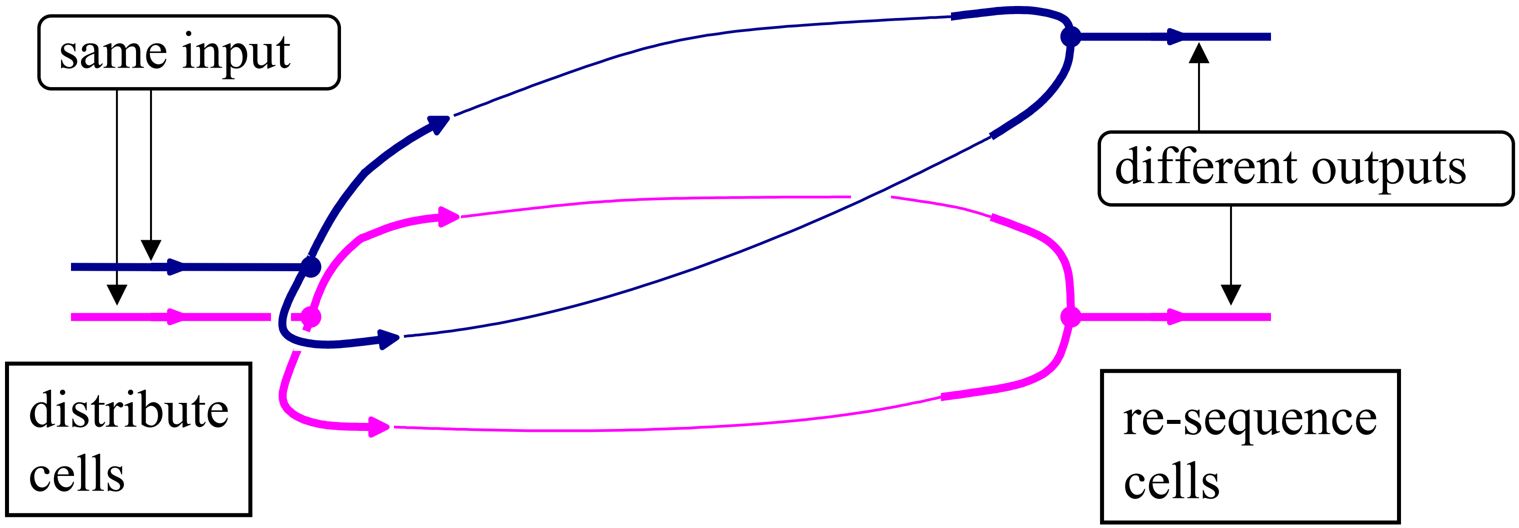
⇓

- Scalable switching fabric architecture:
  - N•log N cost
  - large buffers only on ingress side
- Performance simulation:
  - fully non-blocking
  - delay within 20-60 % of ideal output queueing
  - without internal speedup

Cell Distribution Methods

- Aggregate traffic distribution:
  - Randomized routing (no backpressure)
  - Adaptive routing (indiscriminate backpressure)

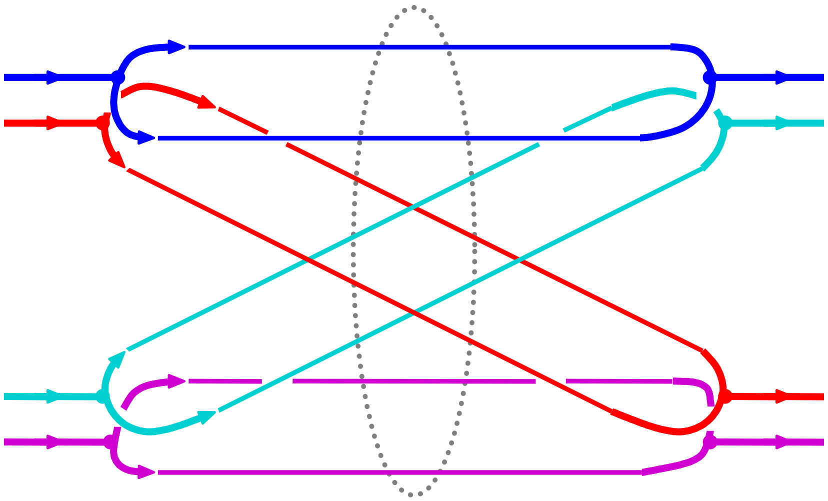
  ⇒ load balancing on the long-term only



- Per-flow traffic distribution:
  - Per-flow round-robin (PerFlowRR)
  - Per-flow imbalance up to 1 cell (PerFlowIC)

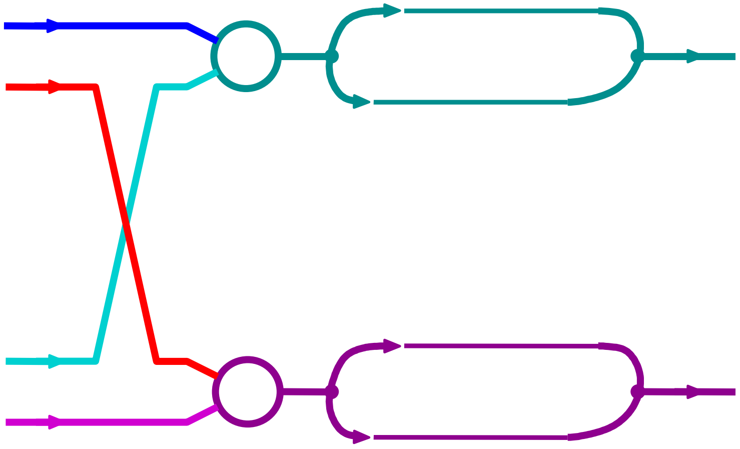
  ⇒ accurate load balancing, on a shorter-term basis

Too many Flows



- N^2 per chip in the middle stage

Per-output Flow Merging



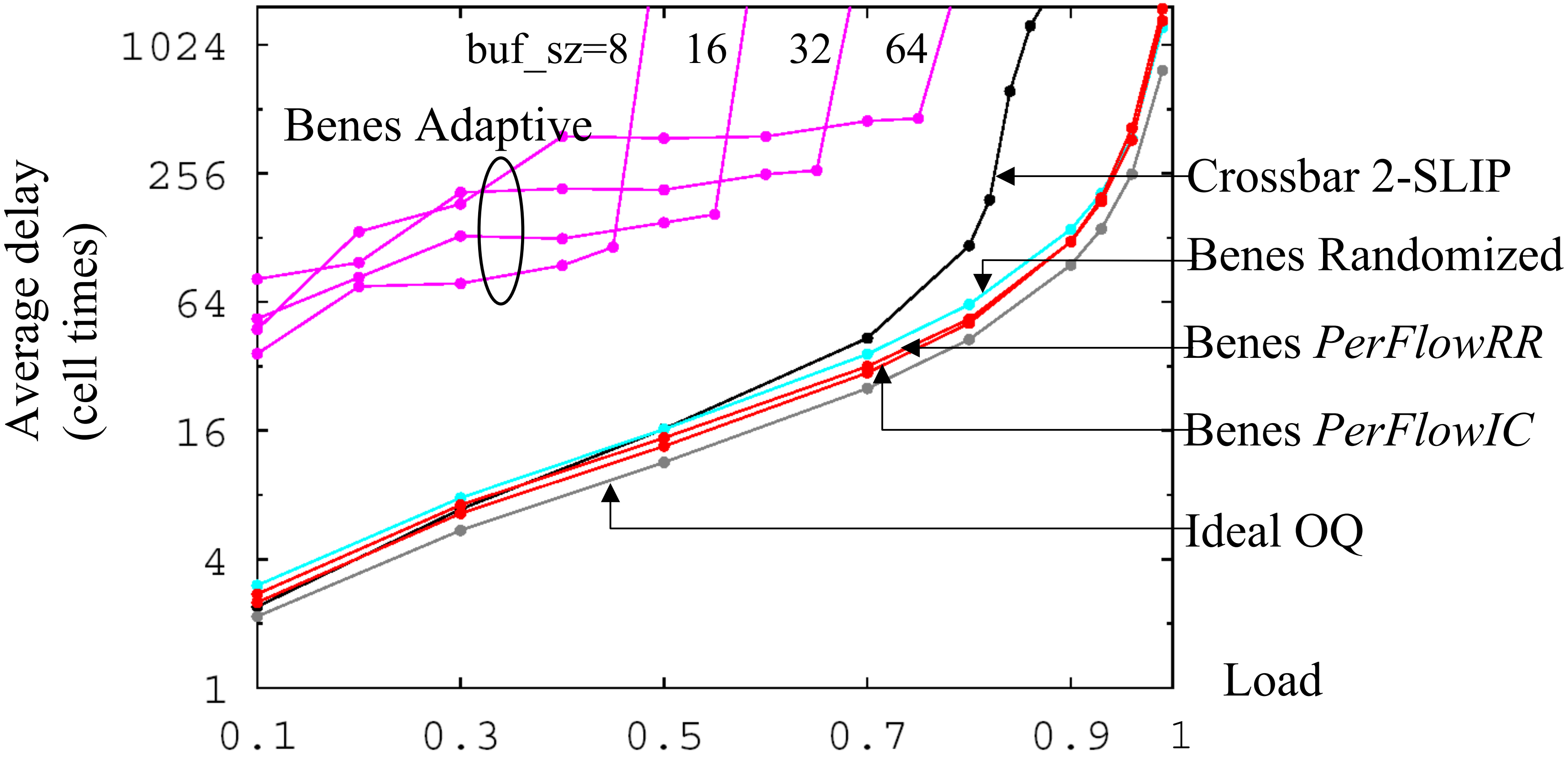
- Retains the benefits of per-flow backpressure
- N flows per link, everywhere

  - Re-sequencing needs to consider flows as they were before merging
  - Freedom from deadlock

Evaluation by Simulation

- Simulation model for the Benes fabric:
  - all link rates = 1 (no speedup)
  - 64 × 64 fabric (or 256 × 256) made of 4 × 4 switches.
  - RTT = 1 cell time (one stage to the next).
  - buffer size = 1 to 3 cells per-flow.
  - report only queueing delay.
- To verify freedom from internal blocking:
  - random permutations.

Bursty/12 Arrivals - Uniform Destinations



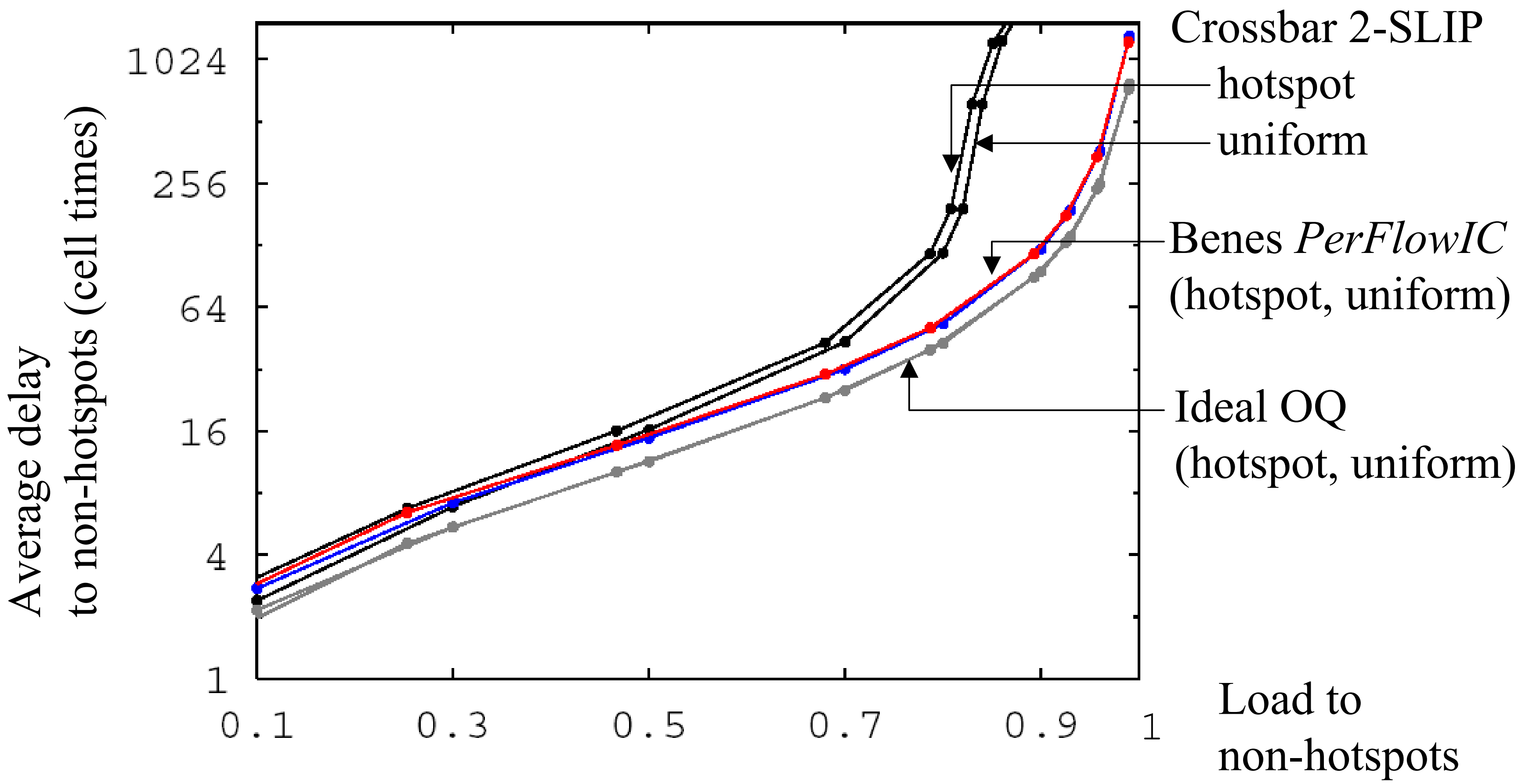
Buffers per Chip:

- Adaptive64: 512 cells / chip
- Randomized: very large buffers
  (no backpressure)
  (16000 cells for 99 % load)
- *PerFlowRR*: 512 cells / chip

Delay:

Benes with per-flow backpressure comes within 20% to 60% of ideal output queueing.

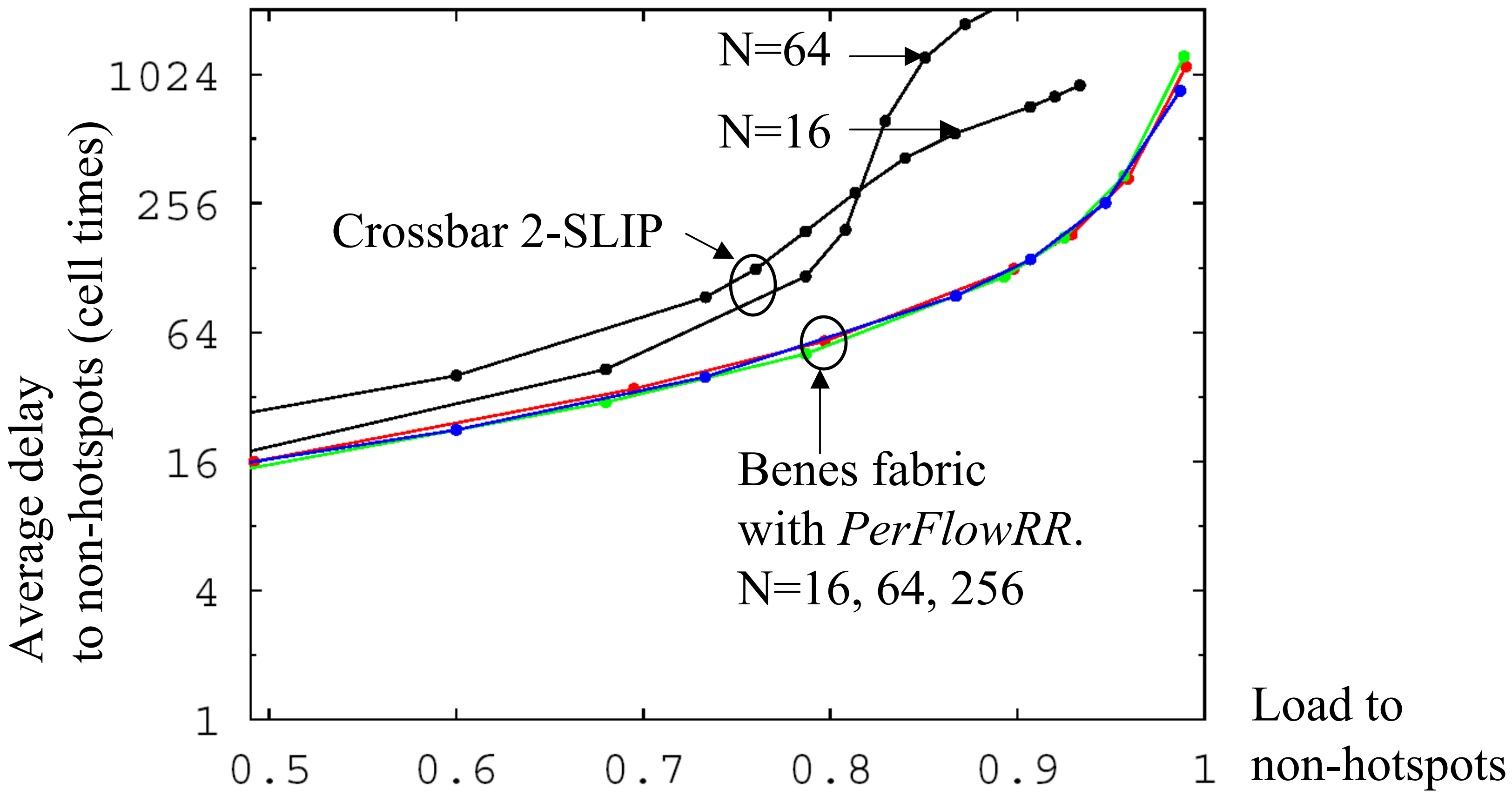
Bursty/12 Arrivals – Hotspots/4



- ➢ 4 out of 64 destinations are hotspots.
- ✓ For the Benes fabric, average delay remain virtually unaffected

  ⇒ Very good flow isolation.

Fabric Size



- ➢ Traffic with bursty/12 arrivals and hotspot/4 destinations.
- ✓ For the Benes fabric, average delay remains virtually unaffected.

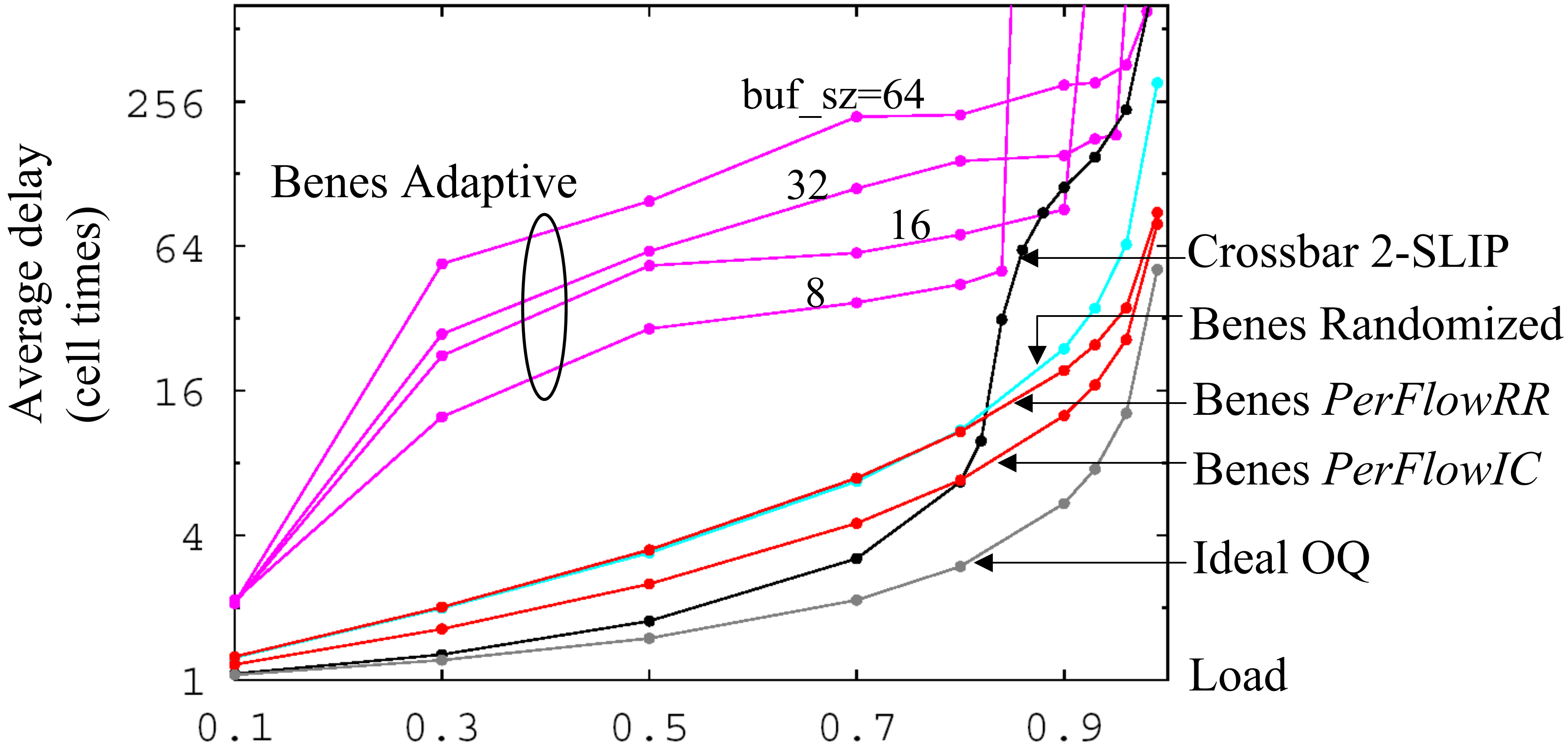
Summary:
Benes Fabric with Internal Backpressure

- Multi-path routing & re-sequencing + per-flow backpressure.
- Per-output flow merging for O(N) switch cost.

⇒ Scalable switching:

- O(N•log N)
- large buffers only on ingress line cards
- freedom from deadlock
- no speedup needed
- fully non-blocking
- performance very close to ideal OQ

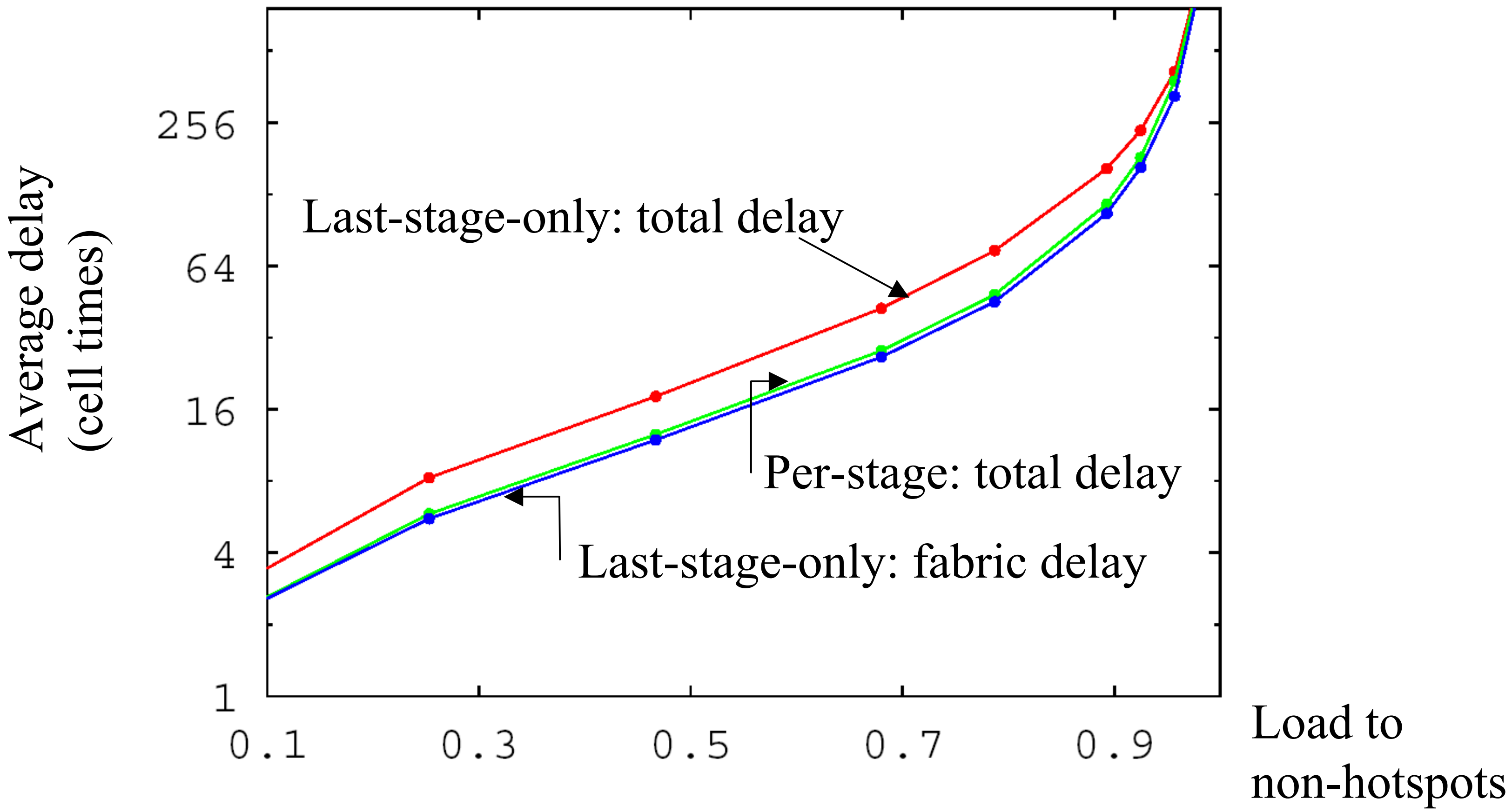
Smooth Arrivals - Uniform Destinations



✓ Randomized cell distribution requires buffer sizes from 5 to 450 cells.

✓ *PerFlowIC* yields 30% to 60% lower delay than *PerFlowRR*.

Alternative Cell Re-Sequencing Methods



- ➢ Traffic with bursty/12 arrivals and hotspot/4 destinations.
- ✓ Per-stage re-sequencing is strictly better than last-stage-only re-sequencing both in terms of implementation cost and in terms of performance.